Reading in a Foreign Language April 2018, Volume 30, No. 1
ISSN 1539-0578 pp. 92–107

The threshold of anxiety in low-stakes testing for foreign language reading

Hitoshi Mikami
Chubu University
Japan

Chi Yui Leung
Nagoya Gakuin University
Japan

Lisa Yoshikawa
Hiroshima University
Japan

Abstract

The following question has yet to be answered by researchers: when does anxiety have a systematic downward bias on foreign language (FL) reading test scores? The results of the correlation and regression analyses conducted in this study indicate that, at least in the case of university-level English learners in Japan, anxiety-induced bias occurs in a low-stakes FL reading test when the test becomes objectively challenging for students. Our data also suggest that concerns about unsuccessful text comprehension play a central role in the elevation of anxiety in a low-stakes test situation.

Keywords: anxiety, perceived language competence, test performance, low-stakes test

Anxiety arousal in foreign language (FL) tests is a curious thing: sometimes we just complete a test without noticing that we have ever been nervous during test taking; meanwhile, we get butterflies in our stomach on other occasions and concentrating upon problem-solving becomes extremely difficult. Both scenarios would be familiar to most FL learners, yet seem utterly different in phenomenology. The aim of this study is to provide a new perspective on such differences in anxiety arousal in low-stakes FL reading (FLR) tests (i.e., tests that have little or no negative consequences for the students taking them). In the realm of FLR research, little discussion has been conducted thus far on when (or in what situation) anxiety has a systematic downward bias on FLR test scores. In this article, we argue that anxiety-induced downward bias occurs in a low-stakes FLR test when FL learners confront a test in which they feel under pressure to exert intense cognitive effort in order to demonstrate successful performance. As the test is perceived as a threat in this situation, anxiety about FLR is intensified; consequently, anxiety compromises test performance. To further explain the background of this study and the

foundations of our argument, the following section will review the relevant literature that led to the current study.

Literature Review

Test Anxiety in Language Assessment

If we take the position of the ability approach (Chalhoub-Deville & Deville, 2005), the aim of language assessment is to correctly estimate one's linguistic competence based on his/her performance in a set of language activities, i.e., tests and tasks (here, we regard tests and tasks as the same entity in terms of psychological testing). In this paradigm, as an example, one's score on an FLR test should be the best reflection of his/her FLR competence.

As our own experience tells us, however, evaluative events sometimes trigger anxiety arousal. Anxiety that we experience in test-taking situations is called test anxiety. That is, the set of emotional, cognitive, and somatic responses that occurs together with the worry about possible negative outcomes on evaluative events (Spielberger, 2010). Test anxiety has been treated as the "major source of construct-irrelevant systematic variance in test scores" (Zeidner, 2010, p.1766) because excessive anxiety arousal in a test-taking situation is consequential to (a) less successful information storage and processing (Moran, 2016), and (b) frequent occurrence of interfering thoughts (Sellers, 2000). Indeed, a large body of psychological studies documented this anxiety–performance interface in high-stakes tests (i.e., tests whose scores form the basis of high-stakes decision making) (e.g., Bellinger & Decaro, 2015; Haladyna & Downing, 2005). Also, when we narrow down our focus to FL assessment, anxiety is even associated with poorer performance in low-stakes tests (Huang & Hung, 2013; MacIntyre, Noels, & Clément, 1997; Madsen, 1982). What is important here is the fact that FL learning research data has predominantly been collected (a) through tests or (b) during test taking. An example of the former is observed in situations where we conduct tests in pre- and post-treatment stages so as to confirm the development (or lack thereof) of FL proficiency for either research or educational purposes. Also, when researchers are interested in the processing mechanism of FL, a test plus think-aloud, eye-tracking, and brain imaging are employed. To use a metaphor, then, anxiety in FL learning research is like the sword of Damocles: the inclusion or creation of individuals who are anxious about tests in a data collection process increases the possibility of making erroneous conclusions.

Anxiety Specific to FLR

As tests are at the core of FL assessment, the link between anxiety and FL performance has been of consistent interest in the realm of FL learning research (see, e.g., Jeong et al., 2016; Scovel, 1976); however, as Güvendir (2014) has pointed out, the amount of literature on the anxiety–performance interface in the domain of FLR is somehow limited as of yet.

Saito, Garza, and Horwitz (1999) conducted, to our knowledge, the first study that differentiated the anxiety specific to the reading of FL (FLR anxiety) from anxiety about general FL use (general FL anxiety). In their study, approximately 59% of the variance was not shared between the measures of general and FLR anxiety. This result was replicated by Sellers (2000) (here, non-

shared variance was 51%) (see also Joo & Damron, 2015) [1]. Sellers (2000) also succeeded in capturing the functional difference between general and FLR anxiety. Compared to their less anxious counterparts, the participants who were high in FLR anxiety showed poorer comprehension of the essential ideas of a given text in a post-reading recall task; meanwhile, the negative impact of FL classroom anxiety (i.e., the measure of general FL anxiety) was limited to the comprehension of supporting ideas. These findings established the concept of FLR anxiety and clarified its unique role in the weaker performance in FLR tests (see also Leung, Mikami, & Yoshikawa, 2017).

The Source of FLR Anxiety

There have also been some, albeit limited, studies that have mentioned or explored the possible sources of FLR anxiety. Madsen (1982) argues that the complexity and difficulty of a test are key to the generation of anxiety in the reading of FL. Saito et al. (1999) confirmed this claim: in their study, students became more anxious about FLR as their perceived difficulty of a reading material increased. Judging from Saito et al.'s (1999) data, the following factors are closely related to the inflation of FLR anxiety: (a) unfamiliar linguistic features (e.g., vocabulary and grammar) in a text, (b) the incomplete comprehension of a text, and (c) the text length. The influence of these factors on anxiety arousal were also documented in later studies (Bektaş-Çetinkaya, 2011; Güvendir, 2014; Joo & Damron, 2015; Zhou, 2017).

Güvendir (2014) also added a test-taking situation to the possible source of anxiety arousal in FLR. During their self-paced FL reading, Güvendir's participants (N = 30) self-reported that they would become nervous provided that they were dealing with the same reading material in an exam (73%) or time-limited (56%) situation. Brantmeier's (2005) data suggested a similar possibility: although FLR anxiety in her study did not have a meaningful relationship with the performance measures, her participants' (advanced FL learners) FLR anxiety increased in a test situation compared to reading alone.

The Manifestation of FLR Anxiety: A Research Gap

As shown in Brantmeier 's (2005) results, despite the seemingly obvious relationship between anxiety and reading performance, not all FLR test scores are proven to be negatively biased by anxeity. Indeed, while the relationship between FLR anxiety and reading performance was documented as an inverse one in the majority of experimental and educational settings (MacIntyre et al., 1997; Saito et al., 1999; Sellers, 2000; Zhao, Guo, & Dynia, 2013), in several cases this relationship was, at least in part, not statistically conclusive (e.g., Brantmeier, 2005; Zhao et al., 2013) or even reported as a positive relationship (Joo & Damron, 2015) (r = .48). These mixed results indicate the possibility of an as yet unconsidered factor changing the nature of the anxiety–performance relationship; this thus raises the question of when (or in what situation) FLR anxeity has a downward bais on FLR test scores. In this article, we argue that anxiety-induced negative bias in FLR tests emerges when tests require intense cognitive effort, and such situations occur when tests become objectively challenging to the students taking them.

If an FL test appeared easy to test takers, the possible negative outcomes on that test event would not be of major concern to them; consequently, there would be little or no increase in their level

of anxiety. On the contrary, when test takers confront a test that matches or exceeds their perceived FL competence (i.e., what test takers believe they can do in the FL), their perceived pressure to make intense cognitive effort towards successful test completion will increase. This time, it is not surprising that test takers' perceived difficulty in a given FL test reaches its threshold and manifests itself as disruptive anxiety. Perceived FL competence in this scenario has a marked impact on fluctuating levels of FL anxiety, and such a relationship was indicated by the results of MacIntyre et al. (1997). In their research, the bias in perceived FL competence, defined as the gap between self-estimated and test-derived FL proficiency scores, was strongly relevant to the level of anxiety that their participants (N = 37) felt in the use of FL. MacIntyre and his colleagues first found that those learners with downward bias in their FL speaking, writing, and listening competence tended to score high on an FL anxiety measure while those with upward bias did the opposite (see also Bonaccio & Reeve, 2010; Mikami, Leung, & Yoshikawa, 2016). Moreover, the anxiety measure used in their study had large inverse correlations with the quality measure of speaking (r = -.55) and writing (r = -.51), and the number of correctly understood passages expressed in a post-listening test (r = -.54). These results highlighted variance in perceived FL competence as the key factor behind the manifestation of disruptive anxiety in the use of FL.

In the same research, however, learners' perceived competence in FLR showed the highest resilience to anxiety. What is more interesting is that the anxiety measure still explained 35% of the variance in performance in text comprehension (measured by the amount of correct understanding of text expressed in a post-reading test) ($r^2 = -.59^2 = .35$). Here, we argue that the following rationale ties together all the results of MacIntyre et al. (1997): first, FL learners are generally capable of estimating their FLR competence with precision; second, this unique nature in turn creates the threshold of FLR anxiety at the point where subjective and objective test difficulty overlap; and third, MacIntyre and his colleagues observed the anxiety–performance interface because, for their participants, the reading material was sufficiently difficult to be perceived as a threat. This rationale also offers a simple explanation for the fact that the inverse relationship between FLR anxiety and reading performance was not observed in several prior studies. That is, when a given test is insufficiently difficult, and thus to be perceived as threatening, the anxiety—performance interface will not be systematic enough to be clearly monitored (Brantmeier, 2005) or may even appear to be a positive relationship (Joo & Damron, 2015).

The Current Research

If the aforementioned rationale is tenable, it will provide future studies with a platform for discussing the relative influence of anxiety on FLR performance. The first goal of this study becomes, therefore, the confirmation of our hypothesis in a low-stakes test situation (i.e., the threshold of FLR anxiety in low-stakes FLR tests exists at the point where subjective and objective test difficulty overlap). We limit our focus to the low-stakes test here because this test situation accords with those used in most prior FLR anxiety studies. As low-stakes tests are the prevalent way of measuring FLR performance, it is also necessary to clarify the amount of anxiety-induced bias generally included in the score of low-stakes FLR tests. Guided by these research interests, we set the following research questions:

RQ1. Does the threshold of FLR anxiety in a low-stakes FLR test exist at the point where subjective and objective test difficulty overlap?
RQ2. When manifested, to what extent does FLR anxiety explain variance in low-stakes FLR test scores?

Methodology

The following three types of data were collected to begin testing our hypothesis: (a) FL learners' self-assessment of FLR anxiety; (b) their performance in a marginal, but non-demanding, low-stakes FLR test and also in a challenging one; and (c) their latest score in a standardized FL test to indicate the relative difficulty of our test measures to each of the learners. If FL learners' assessment of their own FLR competence were generally bias-free, as we hypothesized above, then such a characteristic would be reflected in the score of the FLR anxiety measure. Furthermore, of the two tests, only the challenging one would be perceived as a threat. In this situation, a clear anxiety—performance interface only occurs between the measures for anxiety and for the challenging test. Whether or not the above rationales stand up to scrutiny can be confirmed by correlation analysis. Moreover, provided that our hypothesis is tenable, the implementation of regression analysis allows us to answer research question two, regarding the impact of manifested anxiety on FLR performance.

General Procedure

A total of 69 university students agreed to take part in this study. These students first submitted their current TOEFL ITP® scores and then responded to a questionnaire. In the questionnaire survey, the students first provided their background information (e.g., their age, gender, and language learning history) and then estimated their level of FLR anxiety.[2] The estimate was made without the students being informed of the upcoming activity; that is, the two FLR comprehension tests they would sit. The nature of both tests can be portrayed as low-stakes because the students' test performance affected neither (a) their school record nor (b) the amount of remuneration. Those who completed both the questionnaire and reading tests received 2,000 Japanese yen for their cooperation.

Participants

All 69 recruits were students at a Japanese university and were from 10 different departments. English was their FL in the sense that: (a) they did not speak English as their first or as a heritage language; (b) they had been learning English as one of their school subjects; and (c) they were non-English majors. Of the 69 students surveyed, 63 (91.30%) completed all components of the study described above and thus became the initial sample for this study. Further detailed information on the final sample will be provided in the data screening and analysis section.

Indexes

TOEFL ITP®. The score of TOEFL ITP® indicates each participant's objective English proficiency at the time of investigation.

FLR anxiety. This index reflects the level of anxiety that one feels in FLR activities (k = 4). Each respondent reported on the degree of anxiety arousal that they generally experience in: (a) the reading of FL; (b) evaluative reading tests; (c) the reading of lengthy FL texts; and (d) situations where they are not able to comprehend FL texts (one item for each). Question items were excerpted from Brantmeier (2005) and, as with the original index, all items were answered on a 6-point Likert scale (1 = strongly disagree, 6 = strongly agree). This index was chosen for our investigation because Brantmeier's (2005) study specifically failed to find a meaningful relationship between FLR anxiety and the performance measure. Our argument is, however, that a meaningful relationship will emerge when the difficulty of reading tests is properly controlled. Question items were translated into Japanese and the target language was changed to English where necessary. The wording was refined with the help of five Japanese university students who did not appear in the above data collection. The actual question items employed in this study and their English translations are cited in the appendix (see the Appendix). Despite the small number of question items, the internal consistency of FLR anxiety was adequately high (this point will be considered in the data screening and analysis section).

*FLR comprehension tests***.** Two reading tests were used in this study. Both tests were invented by the Edinburgh Project on Extensive Reading (EPER) (Hill, 1992) and the performance in each test reflects one's comprehension of an English text (Davies & Irvine, 1996; Yamashita, 2008). The essential difference between the two tests (EPERTs) lies in their relative difficulty: one test—EPERT-C—is designed to be easier than the other—EPERT-B (Hill, 1997). EPERT-C consists of a relatively long narrative story (1,408 words) titled "The Book Shop" and 20 questions on the text (k = 20). EPERT-B comprises a 2,027-word narrative story titled "Strange Landlady" and 20 questions on the text (k = 20). In each test, the students were requested to complete the whole test within a span of 30 minutes. Nineteen out of 20 items in EPERT-C and all 20 items in EPERT-B were closed questions. Here, the students filled in incomplete sentences using the information provided in the text (e.g., "*The boy was crying because he* _____").[3] The students could use either English or Japanese for their answers; the variance in FL writing skill thus had little impact on scores (cf., Sellers, 2000, p. 514). The remaining one question in EPERT-C was a multiple-choice question. Here, the participants were given four alternative sentences and required to choose the one that corresponded to the information given in the text.

Data Screening and Analysis

To test our hypothsis, we needed to ensure that EPERT-C functioned as a marginal but non-demanding test while EPERT-B served as a challenging test for all of the students. To that end, we first checked the relative difficulty of the two tests for each stundent. The equivalency table provided in Hill (1997) allowed us to link the EPER levels with the students' current TOEFL ITP® scores (see also Davies & Irvine, 1996; Kanamaru & Educational Testing Service, 2012). Judging from Hill's (1997) table, the B-levels in the EPER criteria (i.e., equivalent to TOEFL

ITP® score range of 480–529) met the requirements for testing our hypotheses. For these individuals, EPERT-B would have been perceived as a threat because of the intense cognitive effort required for successful test completion (see Table 1). We therefore expect this test measure to show a systematic inverse correlation with the anxiety measure. Meanwhile, the B-level students would have retained a greater sense of control over EPERT-C as their FL proficiency was beyond its target level (equivalent to TOEFL ITP® score range of 450–479, see Table 1). EPERT-C is thus expected to show non-systematic correlation with the anxiety measure.

Table 1. *Basic information on the EPERTs*

Test	Title	Text type	Length	EPER level (= TOEFL ITP®)
EPERT-C	The Book Shop	Narrative story	1,408 words	C (= 450–479)
EPERT-B	Strange Landlady	Narrative story	2,027 words	B (= 480–529)

Following this line of reasoning, we first portioned out the 36 students who fell into the B-level range (i.e., 480–529 ± 1SEM on TOEFL ITP ®) (Educational Testing Service, 2016) from the initial sample.[4, 5] We then computed each student's residual score on EPERTs (i.e., the T-score of EPERT-C minus that of EPERT-B). For example, if one's T-score is 60 in EPERT-C and 50 in EPERT-B, this student's residual score is 10. Of the 36 residual scores, one score (= -32.31) was judged to be an outlier (Grubbs' test: $z = 2.69$, $p < .05$), and the data of this student was thus removed from the final sample. The above screening left the data of 35 students for statistical testing. The mean TOEFL ITP® score of these 35 students was 514.12 (*SD* = 16.12, *Skew* = - 0.14, *Kurt* = - 0.46). Their average age was 20.00 (*Mdn* = 20.00; *SD* = 1.68) and the male/female ratio was 0.80. All reported statistical information in the following section was computed based the data of these 35 students (*N* =35). It should be noted that all 35 students spoke Japanese as their first language (L1), and for this reason the data reported in what follows strongly reflect the tendency of university-level Japanese L1 English learners.

The internal consistency of the anxiety index was adequately high (Cronbach's alpha = .82); the arithmetic mean of one's points on all four question items thus became each individual's score on FLR anxiety (score range = 1.00–6.00) (the descriptive statistics on each question item are cited in the Appendix). The Cronbach's alpha of both reading tests was also adequate (EPERT-C, *a* = .76, EPERT-B, *a* = .79). Each correct answer was first converted into the prescribed score points, and the sum of the points attained in a given test became one's test score. The test score ranged from 0.00 to 29.00 (EPERT-C) and 0.00 to 30.00 (EPERT-B).

The skewness of FLR anxiety, EPERT-C, and EPERT-B was 0.07, -0.32, and -0.25, respectively; their standard error of skewness (SES) was 0.40. The skewness/SES ratio of the three indexes ranged, therefore, within ±1.00. As such a range has been considered acceptable for parametric testing, we first computed the Pearson's correlations and then ran a single regression analysis. The statistical programs *R* version 3.1.1 and *G*power 3* (Faul, Erdfelder, Lang, & Buchner, 2007) were employed for statistical computations. Alpha was set at .05 in this study—all reported significances in the results section were corrected for the false discovery rate using the BH method (Benjamini & Hochberg, 2000).

Results

Descriptive Statistics

Table 2 shows the descriptive statistics on the anxiety index, EPERT-C (the marginal but non-demanding test), and EPERT-B (the challenging test) (N = 35, each). On average, the 35 students showed a mildly-high level of FLR anxiety: M = 3.91 out of 6.00 (95% CI [3.67, 4.24]). With regard to test performance, the students as a whole performed better in EPERT-C than EPERT-B (M = 19.49 vs. 17.00) with smaller variance in their test score (SD = 5.06 vs. 6.56, $Kurt$ = -0.62 vs. -1.23).

Table 2. *Descriptive statistics on the three indexes (raw scores)*

Index	*M*	95% CI	*SD*	*Skew*	*Kurt*	α
FLR Anxiety	3.91	[3.67, 4.24]	0.72	0.06	0.04	.82
EPERT-C	19.49	[17.78, 21.20]	5.06	-0.30	-0.62	.76
EPERT-B	17.00	[14.78, 19.22]	6.56	-0.23	-1.23	.79

Correlation

Table 3 shows the basic association (r and its 95% CI) between FLR anxiety, EPERT-C (the marginal but non-demanding test), and EPERT-B (the challenging test) (df = 34, 2-tailed test, each). First, the link between FLR anxiety and EPERT-B was meaningful in terms of its p-value, statistical power, and 95% CI (r = -.50 [95% CI = -.71, -.20], p = .002, $1 - \beta$ = .88. Second, such a link disappeared when the performance index was replaced with EPERT-C: r = -.12 [95% CI = -.44, .22], p = .487, $1 - \beta$ = .10. Lastly, the correlation between the two FLR tests were also meaningful in terms of p-value and 95% CI, but its statistical power did not reach .80 (r = .41 [95% CI = .09, .65], p = .016, $1 - \beta$ = .70)

Table 3. *Correlation matrix (observed variables): r and its 95% CI*

Index	EPERT-C	EPERT-B
FLR Anxiety	- .12 [- .44, .22]	- .50 [-.71, -.20]**
EPERT-C		.41 [.09,-.65]*

Note. df = 35, * = p < .05, ** = p < .01 (2-tailed).

Regression Analysis

The shared variance between FLR anxiety and EPERT-C (the marginal but non-demanding test) was a mere 1% ($r^2 = -.12^2 = .01$). It is clear then, that the regression model that includes EPERT-C will be underpowered. We therefore built only one regression model: the anxiety index was entered as the independent variable and EPERT-B became the dependent variable. The regression model was meaningful in terms of p-value and statistical power (R^2 = .25 [95% CI = .02, .48], $F_{(1, 33)} = 10.82$, p < .001; 1- β = .92). To put it another way, 25% of the variance in

the EPERT-B score was explained by the anxiety measure with a 92% certainty. Also, the effect size (R^2) of this model did not include zero at the 95% CI level.

Discussion

The Threshold of FLR Anxiety in Low-Stakes Tests

The first objective of this study was to confirm the credibility of our hypothesis. That is, that the threshold of FLR anxiety in low-stakes FLR tests can be found at the point where subjective and objective test difficulty overlap. To this end, we created a situation in which the 35 students sat two FLR tests at different difficulty levels—a marginal but non-demanding test and a challenging one. Our prediction was that, under these circumstances, FLR anxiety's association with the test measures becomes clear only in the relation to the challenging test.

First of all, the 35 students' mean score on the anxiety measure reached almost four out of six (M = 3.91, 95% CI [3.67, 4.24]). As stated above, this self-assessment measures the degree of anxiety arousal that the students generally experience in FLR. It is fair to say, then, that our 35 participants were aware of their mild tendency to become anxious in FLR activities.

Despite such a tendency, however, their performance in EPERT-C (the marginal but non-demanding test) did not show a systematic association with FLR anxiety (r = -.12 [95% CI = -.44, .22], p = .487, $1 - \beta$ = .10). According to a-priori power analysis, the sample size needed to attain in order to yield $1 - \beta \geq 80$ for this correlation is $N \geq 154$. This situation changed as we replaced the performance index with EPERT-B (the challenging test), r = -.50 [95% CI = -.71, -.20], p = .002, $1 - \beta$ = .88. This time, despite our relatively small sample size, the correlation fulfills the requirement of meaningful correlation in the field of FL affective research (i.e., ± .30 in the point estimate) (Dörnyei & Ushioda, 2011) and the certainty of correlation reached 88%.

Although we must be cautious in the interpretation of underpowered correlation, we argue that such a correlational change supports our claim. We know that anxiety about FLR elevates when learners are dealing with subjectively challenging tests (Saito et al., 1999). If our students had a downward bias in their perceived FLR competence, then such an underestimation was in turn reflected in their anxiety estimate (MacIntyre et al., 1997) and both test measures would have been associated with the anxiety index (see Table 4). In this study, EPERT-C (equivalent to TOEFL ITP® score range of 450–479) was a marginal but non-demanding test for our 35 B-level learners (equivalent to TOEFL ITP® score range of 480–529). The difficulty of EPERT-C was, in that sense, high enough to be associated with the anxiety index provided a downward bias existed in the students' estimation on FLR competence. On the other hand, if the 35 students as a group had an upward bias in their perceived FLR competence, the anxiety index would not have been correlated with EPERT-B because, in this case, the performance measure with a meaningful correlation to FLR anxiety would be limited to tests beyond learners' current FL proficiency (see Table 4). If the students were generally capable of estimating their FLR competence with precision, then the anxiety–performance interface would appear only between the anxiety index and EPERT-B—the objectively challenging test (see Table 4). This is what we observed in this study, and therefore, it is safe to say that our hypothesis—that the threshold of FLR anxiety in

low-stakes FLR tests exists at the point where subjective and objective test difficulty overlap—is acceptable, at least for Japanese L1 English learners.

Table 4. *Direction of bias in perceived FLR competence and expected correlation patterns*

Direction of Bias	*r* (Anxiety–EPERT-C)	*r* (Anxiety–EPERT-B)
Downward	Y	Y
Upward	N	N
Bias-free	N	Y

Note. Y = $p \leq .05$ and $1 - \beta \geq .80$; N = $p > .05$ or $1 - \beta < .80$.

As for the correlation between the two FLR tests, we yielded a significant *p*-value and the correlation was non-zero at the 95% CI level; however, the certainty of our correlation was 70% and this is 10% lower than the ideal level (Cohen, 1992). That is to say, there was an inconclusive tendency for those who performed better in EPERT-C to do the same in EPERT-B. This result is not particularly surprising given that all 35 students' FL proficiency exceeded the target level of EPERT-C, while EPERT-B was challenging for all of them. This can be confirmed in Table 2: the 35 students performed better in EPERT-C than EPERT-B on average (*M* = 19.49 vs. 17.00) with small variance in their test score (*SD* = 5.06 vs. 6.56, *Kurt* = -0.62 vs. -1.23). The inconclusive correlation between two test measures can thus be attributed to our sampling system.

The Relative Impact of Anxiety on Low-Stakes FLR Test Scores

The second research question concerns the impact of manifested FLR anxiety on the performance in low-stakes FLR tests. The regression model suggests the answer to this topic: when manifested, 25% of the variance in the score of a challenging low-stakes FLR test (i.e., EPERT-B) was explained by anxiety. This result first accords with the finding of prior studies in terms of the negative influence of anxiety on the reading of FL (MacIntyre et al., 1997; Saito et al., 1999; Sellers, 2000; Zhao et al., 2013).

The above results also added a new explanation as to why prior studies, such as that of Brantmeier (2005), did not observe the influence of anxiety on the performance measure: when our measurement reflects one's general propensity to become anxious in FLR, its negative influence will appear in relation to the objectively challenging tests. In addition, when we once again focus on the correlation between FLR anxiety and EPERT-C (the marginal but non-demanding test) (*r* = - .12 [- .44, .22]), the upper bound of 95% CI indicates that the positive correlation between FLR anxiety and the performance of non-demanding FLR tests is a realistic possibility (cf., Joo & Damron, 2015). This result also supports the value of taking test difficulty (as perceived by our participants) into consideration when researching FLR anxiety.

Another interesting result obtained through regression analysis is that FLR anxiety explained the variance in the score of a low-stakes test. This result clarified the fact that the reasons for anxiety arousal in FLR tests are not limited to apprehension about the loss of academic or economic benefit—the negative consequences typically attached to high-stakes tests (cf., East, 2014). The anxiety generating factors that constituted our anxiety index were the test situations themselves, the reading of FL, the length of FL texts, and the incomplete understanding of FL texts (see

Methodology and the Appendix). The last three factors are more related to reading *per se* and the efficiency of reading, rather than the loss of instrumental benefit. Saito et al. (1999) also considered the efficiency of reading as a sub-construct of FLR anxiety. In particular, they argued that FL learners' propensity to try to understand everything stated in a text ties the encounter with unfamiliar words and grammar with anxiety arousal (see also Bektaş-Çetinkaya, 2011). Following this line of reasoning, confrontation with challenging and lengthy FL texts also raises the level of anxiety by increasing the perceived probability of imperfect reading comprehension (cf., Güvendir, 2014; Zhou, 2017). Furthermore, incomplete understanding of FL texts naturally becomes the focus of apprehension because, when it occurs, one becomes aware of the fact that the best one can achieve in their reading attempt is a less than perfect comprehension. These expectations of failure in intellectual challenges in text comprehension, seems, then, the key in the arousal of FLR anxiety in a low-stakes test situation.

Conclusion

Summary and Implications

The results of this study suggest the existence of negative bias in low-stakes FLR test scores. Such a bias appears when learners confront tests in which they feel apprehensive about unsuccessful reading comprehension. Because perceived FL competence is accurate in the domain of reading, the arousal of such anxiety takes place when a test becomes objectively challenging to learners. The sampling system of this study allows us to conclude that our hypothesis on the threshold of FLR anxiety is acceptable, at least in the case of university-level Japanese L1 English learners.

The downward bias in test scores should be removed from data analysis if, as we do, one wants to assess FLR competence with accuracy. The findings of this article can be used to do so. Here, we propose two different approaches to reducing anxiety-induced bias from our assessment. First, if we limit our focus to statistical testing, we can document the degree of FLR anxiety together with the other data, including the relative difficulty of a given reading test to test takers, and then statistically control the influence of anxiety in data analysis.

The remaining approaches concentrate on the improvement of classroom language assessment. The use of challenging reading tests for monitoring linguistic development is prevalent in language classrooms. If teachers wish to use such tests, it is highly recommended that they foster students' perceived importance of FLR before starting the assessment. In Mikami et al. (2016), the attribute we call "lack of motivation for FLR" showed a strong positive correlation with the anxiety measure used in this study (r = .54). Lack of motivation represents, as its name suggests, a lack of personal interest or self-recognized importance in studying FLR. This result indicates that learners who lack motivation to study FLR are more prone to be anxious about reading FL. In order to overcome this situation, learners need to somehow internalize the value of studying FLR. On this point, Yamashita (2013) reported that the implementation of a 15-week extensive reading program fostered her students' intellectual value in FLR (r = .26) as well as decreasing their anxiety toward FLR (r = .34). These data show the clear benefit of extensive reading on the

inhibition of anxiety in the reading of FL, and thus the use of extensive reading seems a feasible way to deal with the issue of FLR anxiety in our everyday FL assessment.

Directions for Future Studies

Two intriguing questions remain for future studies. First, in this study, one attribute of anxiety arousal in a low-stakes FLR test situation was confirmed for a specific learner group (i.e., university-level Japanese L1 English learners). On this point, Saito et al. (1999) reported that the writing system of the target language (e.g., the dependability of symbols) acts as another factor influencing FL learners' anxiety level in the reading of FL (for a similar discussion, see Joo & Damron, 2015; Zhao et al., 2013). There still is a possibility, then, the threshold for FLR anxiety differs according to the target language of learners. The results of this study should therefore be tested for generalizability (i.e., "Do all FL learners generally have their threshold of anxiety for low-stakes FLR tests at the point where subjective and objective test difficulty overlap?").

Second, it would be interesting for future studies to examine whether the correlation between FLR anxiety and performance in low-stakes FLR tests remains the same or becomes stronger when we use an objectively difficult test as a performance measure. To use the criteria of EPER, a difficult test for this study would be EPERT-A (equivalent to TOEFL ITP® score range of 530–549) or EPERT-X (equivalent to TOEFL ITP® score from 550 or more) (Hill, 1997). For that research purpose, researchers could record the anxiety-relevant symptoms that students develop together with the increase in test difficulty (i.e., state anxiety in FLR), in addition to the general tendency to become anxious in FLR (i.e., trait anxiety in FLR). As shown in this study, accurately perceived FLR competence is at least one indicator of the estimate of trait anxiety in FLR. Where this is the case, the correlation between trait anxiety in FLR and test performance may be fixed once test difficulty reaches learners' current FL proficiency. However, our results also suggest that the level of anxiety in low-stakes FLR tests increases in line with the rise in perceived probability of imperfect text comprehension. The impact of anxiety on performance in FLR tests may therefore show a steady increase if we monitor the changes in learners' state anxiety together with the rise in test difficulty. If future studies confirm the difference in the impact of trait and state anxiety on the performance of a difficult FLR test, this would firstly add further support to the claims made in this study, and also demonstrates appropriate usage of two different anxiety measures.

Last but not least, one limitation of our analysis is the need for replication with larger samples. The regression model used in this study would stand the test of time as the risk of type 1 and 2 errors is smaller than 1% and 8%, respectively, $R^2 = .25$, $F_{(1, 33)} = 10.82$, $p < .001$; $1\text{-}\beta = .92$. There would thus be no problem in future FLR studies using our estimate as the reference value for their investigations. Future replication studies should, however, make the 95% CI of the model more precise by increasing their sample size. Although the non-zero effect size of our model at the 95% CI level [02, .48] showed a high level of confidence in our model, we admit that a more precise estimate of the population parameter would deepen the understanding of the amount of bias included in challenging low-stakes FLR test scores.

The results of this study shed light on the role of test difficulty in the manifestation of FLR anxiety. As our argument is supported by data and has the potential to further clarify the nature

of anxiety in FLR, it would be beneficial for future studies to provide answers to the above-stated open questions and to overcome the limitations of this study.

Acknowledgments

We are grateful to RFL's anonymous reviewers for their informative comments on this manuscript. We would also like to thank Natalie-Anne Hall (The University of Manchester) for English language editing. Note that this work was supported by Chubu University Grant B (29 | L02B), and JSPS KAKENHI Grant Numbers JP16K16887 and JP17K13500.

Notes

1. Joo and Damron (2015) reported $r = .70$ for the correlation between the measure of general FL anxiety and FLR anxiety; thus, their shared variance was 49% ($r^2 = .70^2 = .49$).

2. The students also responded to 34 question items in another section of the questionnaire, because another study was being conducted concurrently.

3. The copyright issue did not allow us to cite the original text nor a question item; the question used in the example is thus a hypothetical question.

4. 1SEM in TOEFL ITP® is ± 14 (Educational Testing Service, 2016).

5. In total, the data of 27 students were excluded from the data analysis: according to the criteria of the EPER, 16 students were at the X-level (equivalent to TOEFL ITP® score from 550 or more), one student was at the A-level (equivalent to TOEFL ITP® score range of 530–549), seven students were at the C-level, and the remaining three were at the D-level (equivalent to TOEFL ITP® score range of 400–449).

References

Bektaş-Çetinkaya, Y. (2011). Foreign language reading anxiety: A Turkish case. *The Journal of Language Teaching and Learning, 1*, 44–56.

Bellinger, D., & Decaro, M. S. (2015). Mindfulness, anxiety, and high-stakes mathematics performance in the laboratory and classroom. *Consciousness and Cognition, 37*, 123–132. doi:10.1016/j.concog.2015.09.001

Benjamini, Y., & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics, 25*, 60–83. doi:10.3102/10769986025001060

Bonaccio, S., & Reeve, C. L. (2010). The nature and relative importance of students ' perceptions of the sources of test anxiety. *Learning and Individual Differences, 20*, 617–625. doi:10.1016/j.lindif.2010.09.007

Brantmeier, C. (2005). Anxiety about L2 reading or L2 reading tasks? A study with advanced language learners. *The Reading Matrix, 5*, 67–85.

Chalhoub-Deville, M., & Deville, C. (2005). Handbook of research in second language teaching and learning. New Jersey: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. doi:10.1037/0033-2909.112.1.155

Davies, A., & Irvine, A. (1996). Comparing test difficulty and text readability in the evaluation of an extensive reading programme. In M. Milanovic & N. Saville (Eds.), *Studies in language testing 3* (pp. 165–183). Cambridge: Cambridge University Press.

Dörnyei, Z., & Ushioda, E. (2011). *Teaching and researching motivation* (2nd ed.). London: Longman.

East, M. (2014). Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform. *Language Testing*, *32*, 1–20. doi:10.1177/0265532214544393

Educational Testing Service. (2016). *TOEFL ITP Test Taker Handbook.* Retrieved from https://www.ets.org/s/toefl_itp/pdf/toefl_itp_test_taker_handbook.pdf

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. doi:10.3758/BF03193146

Güvendir, E. (2014). Using think-aloud protocols to identify factors that cause foreign language reading anxiety. *The Reading Matrix*, *14*(2), 109–118.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27. doi:10.1111/j.1745-3992.2004.tb00149.x

Hill, D. R. (1992). *The EPER guide to organizing programs of extensive reading.* Edinburgh: University of Edinburgh (IALS).

Hill, D. R. (1997). Graded (basal) readers—choosing the best. *The Language Teacher Online*. Retrieved from http://jalt-publications.org/old_tlt/files/97/may/choosing.html

Huang, H.-T. D., & Hung, S.-T. A. (2013). Comparing the effects of test anxiety on independent and integrated speaking test performance. *TESOL Quarterly*, *47*, 244–269. doi:10.1002/tesq.69

Jeong, H., Sugiura, M., Suzuki, W., Sassa, Y., Hiroshi, H., & Ryuta, K. (2016). Neural correlates of second-language communication and the effect of language anxiety. *Neuropsychologia*, *84*, e2–e12. doi:10.1016/j.neuropsychologia.2014.11.013

Joo, K. Y., & Damron, J. (2015). Foreign language reading anxiety: Korean as a foreign language in the United States. *Journal of the National Council of Less Commonly Taught Languages*, *17*, 23–55.

Kanamaru, T., & Educational Testing Service. (2012). *TOEFL® ITP test koushiki test mondai & gakusyuu guide. [The official TOEFL® ITP tests & guide]*. Tokyo: kenkyu-sya.

Leung, C. Y., Mikami, H., & Yoshikawa, L. (2017, September). *Effects of anxiety on word recognition during second language reading: An eye-tracking study*. Paper presented at the 27th conference of the European Second Language Association, Reading, UK. Retrieved from https://www.reading.ac.uk/celm/media/1159/eurosla-conference-abstracts-2017.pdf

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning*, *47*, 265–287. doi:10.1111/0023-8333.81997008

Madsen, H. S. (1982). Determining the debilitative impact of test anxiety. *Language Learning*, *32*, 133–143. doi:10.1111/j.1467-1770.1982.tb00522.x

Mikami, H., Leung, C. Y., & Yoshikawa, L. (2016). Psychological attributes in foreign language reading: An explorative study of Japanese college students. *The Reading Matrix, 16*(1), 49–62.

Moran, T. P. (2016). Anxiety and working memory capacity : A meta-analysis and narrative review. *Psychological Bulletin, 142*, 831–864. doi:10.1037/bul0000051

Saito, Y., Garza, T. J., & Horwitz, E. K. (1999). Foreign language reading anxiety. *The Modern Language Journal, 83*, 202–218. doi:10.1111/0026-7902.00016

Scovel, T. (1976). The effect of affect on foreign language learning: a review of the anxiety research. *Language Learning, 28*, 129–142. doi:10.1111/j.1467-1770.1978.tb00309.x

Sellers, V. D. (2000). Anxiety and reading comprehension in Spanish as a foreign language. *Foreign Language Annals, 33*, 512–520. doi:10.1111/j.1944-9720.2000.tb01995.x

Spielberger, C. D. (2010). Test anxiety inventory. In I. B. Weiner & W. E. Craighead (Eds.), *The Corsini encyclopedia of psychology* (4th ed., p. 1767). New Jersey: Wiley.

Yamashita, J. (2008). Extensive reading and development of different aspects of L2 proficiency. *System, 36*, 661–672. doi:10.1016/j.system.2008.04.003

Yamashita, J. (2013). Effects of extensive reading on reading attitudes in a foreign language. *Reading in a Foreign Language, 25*, 248–263.

Zeidner, M. (2010). Test anxiety. In I. B. Weiner & W. E. Craighead (Eds.), *The Corsini encyclopedia of psychology* (4th ed., pp. 1764–1766). New Jersey: Wiley.

Zhao, A., Guo, Y., & Dynia, J. (2013). Foreign language reading anxiety: Chinese as a foreign language in the United States. *The Modern Language Journal, 97*, 764–778. doi:10.1111/j.1540-4781.2013.12032.x

Zhou, J. (2017). Foreign language reading anxiety in a Chinese as a foreign language context. *Reading in a Foreign Language, 29*, 155–173.

Appendix

Question Items of FLR Anxiety Index and Their Descriptive Statistics

No.	Question (English translation)	*M*	*SD*	*Skew*	*Kurt*
1	英語の文章を読まなければならない時に 不安を感じる。 (I become anxious when I have to read in *English*.)	3.54	0.94	-0.46	0.29
2	英語の長文を読まなければいけないと 不安になる。 (I fear having to read lengthy texts in *English*.)	3.66	1.09	0.06	-0.89
3	英語の長文を読む時に, 内容が理解できないのではないかと心配になる。 (I fear not understanding lengthy *English* texts.)	4.63	0.90	-0.41	0.73
4	英文を読んだ後で，その内容に関する質問に 答えなければいけないと不安になる。 (I become anxious when I have to answer *questions* about what I have read in *English*.)	3.83	0.91	0.84	-0.22

Note. $N = 35$; Score range = 1.00–6.00, each; italics denote sections that were modified from the original descriptions.

About the Authors

Dr. Hitoshi Mikami is a faculty member in the Department of English Language and Culture at Chubu University. His research interests include individual differences in language learning, language learning in study abroad contexts, and plagiarism. E-mail: mikami_h@isc.chubu.ac.jp.

Chi Yui Leung is a lecturer at Nagoya Gakuin University. His research interests include individual differences in reading performances, and eye-movement control in second language reading. E-mail: sieileung@gmail.com.

Lisa Yoshikawa has previously worked as a lecturer at the Institute for Foreign Language Research and Education at Hiroshima University, and is currently a lecturer of the Institute of Liberal Arts and Sciences at Toyohashi University of Technology, Aichi, Japan. She examines the cognitive processes in L2 reading, individual differences in L2 reading comprehension, and assessment of L2 reading ability. Email: yosikawa@las.tut.ac.jp.